Analyzing Quarterbacks in the NFL

Submitted To

Joydeep Ghosh
Cheng Lee

Prepared By

Aaron Alaniz
Iñaqui Delgado
Greg Scaffidi

EE380L
Electrical and Computer Engineering Department
University of Texas at Austin

Spring 2013

CONTENTS

TERMS

1. *QBR - Total Quarterback Rating- A measure of quarterback performance. Similar to Passer Rating, but involves a more sophisticated computation.*
2. *CPCT - Completion Percentage*
3. *QBNC - Quarterback data No Combine*
4. *QBWC - Quarterback data With Combine*
5. *Heisman - "most outstanding player in collegiate football"*
6. *NFL Combine - Set of physical/mental tests that measure a college football player's size, speed, strength, agility, fitness and intelligence. It is administered every February and used by NFL teams when deciding which players to select in the NFL draft.*
7. *Passer Rating - A measure of a quarterback's performance or efficiency. In the NCAA, it is defined as ((8.4 x YDS) + (330 x TD) + (100 X COMP) - (200 x INT)) / ATT*
8. *YDS - Passing Yards*
9. *TD - Passing Touchdowns*
10. *COMP - Pass Completions*
11. *INT - Passing Interceptions*
12. *ATT - Pass Attempts*
13. *AY/ A - Adjusted Passing Yards Per Attempt - (YDS + 20 * TD - 45 * INT) / ATT*
14. *Wonderlic - The intelligence test in the NFL combine*

ABSTRACT

Quarterbacks, at any level, deal with the pressure of being the most valuable player on a football team. In the NFL, large sums of money and powerful sports markets exacerbate this pressure and leave teams constantly searching for the next Joe Montana or Tom Brady. However, the questions still remain: what factors contribute to a quarterback's success in the NFL, and how can they be identified prior to time they play their first snap? This report highlights our work analyzing the performance of a quarterback's first full season as an NFL starter. We selected statistics from each quarterback's collegiate career and used this data to build models based on each quarterback's performance and physical attributes. Furthermore, we used these models for prediction, classification, and clustering. Our findings provide a duality of evaluation. On one hand, our results indicate a high degree of difficulty in identifying which quarterbacks lead to the ultimate successes in the NFL: such as wins, division titles, or Super Bowl rings. In contrast, we concluded that some quarterback statistics at the NFL level can be predicted by college performance.

1.0 INTRODUCTION

Finding a quarterback capable of yielding success at the NFL level proves to be one of the most difficult tasks in all of professional sports. In some cases, investing in a young promising quarterback who ends up gravely underachieving can set back franchises years because of the financial burden they put on a team's budget. In the 1998, NFL scouts deemed quarterbacks Peyton Manning and Ryan Leaf the undisputed top two picks in the NFL draft. Both came with an impressive set of college statistics and finished 2nd and 3rd in the Heisman trophy race [1]. NFL scouts were split on who was actually better between the two but seemed to agree that they would both have a long term positive impact on any team who selected them [2]. The Indianapolis Colts drafted Peyton Manning number one overall and the San Diego Chargers drafted Ryan Leaf immediately after. Peyton Manning signed for a then rookie record contract of $48 million dollars over six years [3]. The San Diego Chargers followed suit by sealing a $31.25 million dollar deal over four years with Ryan Leaf. To this day, Peyton Manning remains one of the most feared quarterbacks in the NFL and is widely argued as the greatest quarterback who ever played the game. Under the leadership of Manning, the Colts made nine straight playoff appearances and won the Super Bowl in 2006 [4]. In stark contrast, Ryan Leaf went on to have a short, abysmal NFL career. Over five seasons with four teams, Leaf accumulated a grand total of four wins and 14 touchdowns. To add perspective, Manning won 13 games and threw 26 touchdowns in just his second season in the NFL [5].

As indicated by the provided anecdote, the process of scouting and drafting quarterbacks in the NFL is highly difficult and can yield disastrous results. Moreover, the value of and importance of the NFL quarterback has not diminished at all since 1998. There is no doubt that knowing that Peyton Manning would be a first ballot hall of fame[1] caliber player is nearly impossible to predict. In addition, success in football is dependent on so many factors outside of the quarterback position. However, it may have been valuable to the San Diego Chargers to know that based on Ryan Leaf's college statistics, he would not have yielded first round quarterback statistics at the pro level and therefore was not worth pursuing at such a high price. Knowing this could have saved the Chargers millions of dollars and they could have pursued a cheaper,

[1] Player who is voted into the NFL Hall of Fame after one round of voting.

later round alternative such as Matt Hasselbeck who went on to have a respectable NFL career with a Super Bowl appearance. This report explores the possibility of finding statistical evidence in a quarterback's college and NFL combine performance to predict some subset of their NFL statistics of their debut season. First we discuss the source of our data and how we aggregated it into our working data set. The following sections cover our statistical approach to analyzing the data and our corresponding results for each experiment. We then conclude our report with our learnings and potential future work in this research topic.

2.0 DATA

In this project, we seek to evaluate the efficiency with which NFL teams draft quarterbacks. The data we use to model a quarterback's performance can roughly be split into two categories: NFL statistics, and college statistics. When drafting a player into the NFL, teams consider a variety of data about that player's performance in college games, as well as that player's physical characteristics and scores in the NFL combine tests.

2.1 GATHERING

A major challenge to our evaluation of how well NFL teams make use of a player's college stats, is the critical issue of obtaining the same data that is available to the scouts, coaches, and executives of NFL teams. To perform our analyses, we must be able to measure a player's career performance both in the NFL, and prior to joining the NFL. The distinction between these two time periods in a player's career can be viewed as a fundamental temporal boundary in our data set. Without access to a complete database of NFL player statistics containing both NFL and college statistics, we were tasked with scraping Internet sources on both sides of this boundary and compiling the resulting data into one common pool.

2.1.1 Sources

We simultaneously evaluated several potential sources of data. One source, that seemed particularly promising at first, was Dbpedia-- a project with the goal of making the contents of Wikipedia accessible from structured query languages, like SPARQL. However, limitations

imposed by Dbpedia's [2] Virtuoso server back-end on the length of time that queries can take, and the lack of an easy way to export the results, convinced us to look elsewhere. Our data was primarily sourced from two websites, one for college stats[3] and another (Pro Football Reference) for NFL stats[4], because these two proved to be the most complete and accessible for the majority of our data.. We obtained the data by writing Python scripts that allow use of the capabilities provided by the Scrapy [5] project to crawl the player pages at the above websites and extract the pertinent data. We supplemented this data with NFL draft statistics,[6] player salary data,[7] and NFL combine data[8]. Wonderlic scores were particularly hard to find for most players and were individually obtained from a variety of online sources.

2.1.2 Process

Scrapy allowed us to launch web crawlers on a list of domains provided as input. Sports Reference and Pro Football Reference served as our two primary targets for scraping since the data was spread out over a list of subdomains in each case. Each site contained a subdomain for each quarterback, but stored in slightly different ways. Pro Football Reference provided a landing page for just quarterbacks that we extracted the links from. Sports Reference did not contain such landing page and only allowed users access to a list of all players sorted by last name and partitioned by each letter in the alphabet. We retrieved all the links for each letter and used them as input for another spider to determine which players were actually quarterbacks. Each player description contains a position field and we just filtered our links to contain only players who were quarterbacks.

With all the links to quarterbacks gathered, we began to scrape our data from the each player's page. Figure 1 shows an example table of Tom Brady from Pro Football Reference.

---

[2] http://dbpedia.org/About
[3] http://www.sports-reference.com/cfb/players/
[4] http://www.pro-football-reference.com/players/
[5] http://scrapy.org/
[6] http://www.drafthistory.com/index.php/positions/qb
[7] http://www.spotrac.com/
[8] http://nflcombineresults.com/nflcombinedata.php

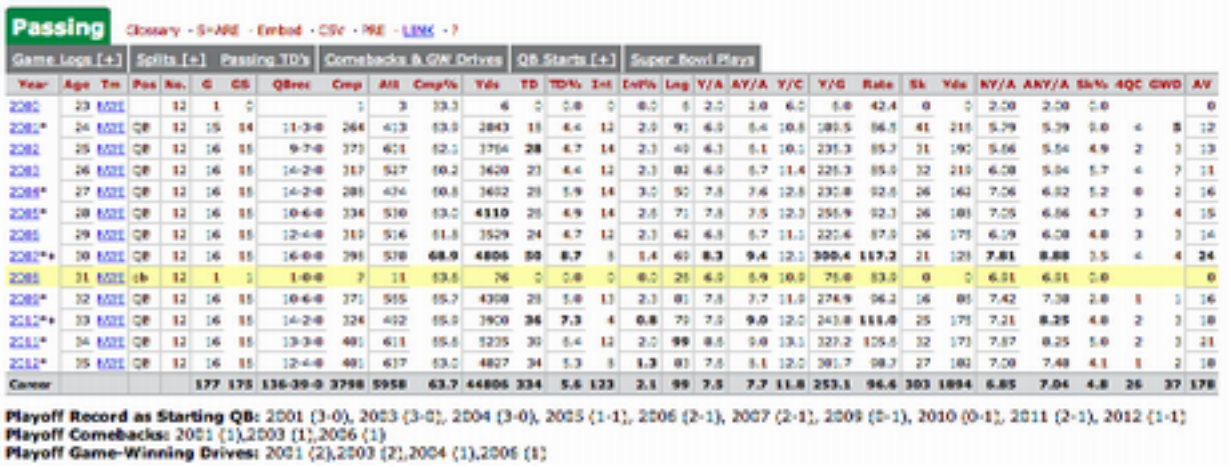

Figure 1. Example Table from Pro Football Reference

We used the first season in which a quarterback started more than seven games to represent their debut NFL season. Revisiting Figure 1, we selected the year 2001 from Tom Brady's table. We were interested in an aggregate form of data at the college level for each quarterback. Figure 2 shows an example table from Sports Reference that contained Tom Brady's college data.

Passing Data Coverage (*=bowl stats included) · Glossary · SHARE · Embed

					Passing							
Year	School	Conf	Class	Pos	Cmp	Att	Pct	Yds	Y/A	AY/A	TD	Int	Rate
1996*	Michigan	Big Ten		QB	3	5	60.0	26	5.2	-3.8	0	1	63.7
1997	Michigan	Big Ten		QB	12	15	80.0	103	6.9	6.9	0	0	137.7
1998*	Michigan	Big Ten		QB	214	350	61.1	2636	7.5	6.8	15	12	131.7
1999	Michigan	Big Ten		QB	180	295	61.0	2217	7.5	7.7	16	6	138.0
Career	**Michigan**				**409**	**665**	**61.5**	**4982**	**7.5**	**7.1**	**31**	**19**	**134.1**

Figure 2. Example Table from Sports Reference

Our web crawler counted the number of years each quarterback played in college and extracted the statistics from the career row on the bottom. We used the number of years to convert some stats into a yearly averages for each player. The code required for each scrape job added up to 153 lines and executed in less than a five minutes.[9]

[9] Source found at http://bit.ly/qb_scraper

2.2 CLEANING

After assembling the scraped data into .CSV files, surprisingly little cleaning was required to begin analysis. However, because our data originated from multiple sources we used the R merge() function to generate a single .CSV to serve as our primary data set. We used the players' names as the criteria for matching the different data together. Our original NFL quarterback data contained 367 players. We also compiled data for 243 college quarterbacks. The resulting merged data set consisted of 240 quarterbacks. The following quarterbacks were lost in the merge: JP Losman, TJ Rubley, AJ Feeley. Looking back through each data set, after performing our analysis, the players above were actually in both files prior to the merge. However, the NFL stats included periods in between the players initials and the names were not matched by the R merge function. This minor oversight should not have had much impact on our analysis overall.

2.2.1 Missing Data

Several of our data factors contained significant missing data. We were missing scores from some of the categories of the quarterback statistics to varying degrees[10]. Out of the 240 quarterbacks that made up our core data set, 214 of them were missing the QBR score, by far the statistic with the most missing data in the NFL data set. If we were to only consider quarterbacks with scores in every one of the categories, we would be left with just 17 quarterbacks to consider. Adding the NFL combine data reduced the size of the data set to just 67 quarterbacks (of all the NFL quarterbacks in our data set, we could only find combine data for 67 of them). Only two quarterbacks had bench press scores available. In this case, if we were to omit all rows with some missing data, we would be left with zero quarterbacks. While including the bench press score creates an extreme example of the problem we faced with missing data, even if we exclude it from the data set and recalculate the number of quarterbacks with a full set of scores, only 12 would remain. Furthermore, visually inspecting the data set reveals that the missing data does not occur randomly. In fact, categories with higher percentages of missing data tend to correlate with other categories that have a similar amount of missing data. The same players tend to be missing statistics in several categories. Because the data isn't missing

[10] http://rpubs.com/sgscaffidi3/NAsInData

at random, our data set is not a very good candidate for mean substitution. In many cases, we are missing data because a particular statistic wasn't being tracked yet (QBR, for instance). We cannot assume that data collected in later decades is representative of quarterbacks in earlier decades. In other cases, the data is missing because it was never meant to be publicly available (e.g. Wonderlic).

2.2.2 Subsetting

A combination of the fact that our data came from several sources and the issues we faced with missing data (significant reductions in size of data set when the data from different sources were merged), we found it useful to consider subsets of the data. For example, we looked at the data with and without the combine scores, and in each case also with and without the draft pick rankings. Some of our data is also highly correlated and entire columns needed to be removed when performing PCA on the college statistics. While subsetting the data made some analyses easier to perform, at times it prevented us from considering the entire breadth of our data over a larger sample of quarterbacks.

2.3 GENERAL ANALYSIS

We performed some preliminary analysis and visualization of our data to reveal basic trends in the predictor variables. This analysis was performed on the data both with and without the combine scores added, because adding the combine scores substantially reduced the number of quarterbacks, as described in section 2.2.1. The R summary() function was extremely useful for reporting the mean, median, minimum and maximum of each variable. Visually inspecting the histogram plots of the data helped to identify cases where binarization and log transformation might be useful in later analysis. Density plots are useful in some case where the range of values that the factor can take is fairly great, but not applicable when the factor has only a few values.

We explored the effects of the reduction in data set size from adding the combine data[11]. The percent change in the minimum, maximum, mean, median and standard deviation for each of the numeric data features is calculated and plotted using barcharts in R. Several of the minimum values in the original data set were zero before adding the combine data, so calculation the

[11] http://rpubs.com/sgscaffidi3/compareCombine

percentage of change after the combine data was added resulted in R reporting the change was infinite (Inf). In some cases, the minimum value did not change, and was zero before and after adding combine data. However, the important takeaway from this exercise is found when comparing the change in mean and standard deviation. The mean values for several popular college statistics (average touchdowns, completions, attempts, and yards) changed by as much as 30 - 40%. The standard deviation for 22 of the 45 features compared changed by more than 10 percent after the combine data was added. These results show that there is likely not enough combine data available to create models that can accurately predict NFL stats based on combine scores.

3.0 APPROACH AND EVALUATION

Our main goal for our models was to determine a relationship between a set of college statistics for each quarterback and their performance in the NFL. Our criterion for data selection left us with a range of 70 to 268 quarterbacks depending on the model or data set we chose to work with. We chose to use linear and logistic regression, principal components analysis (PCA), and clustering for our models. The following sections highlight our methodology, results and evaluation for each model.

3.1 LINEAR REGRESSION

QBNC and QBWC served as the primary data sets of interest for linear regression analysis. The following NFL statistics for each quarterback defined our set of response variables for each model.

- Wins
- Games Started
- Completion Percentage
- Touchdowns
- Interceptions
- Yards

- Sacks
- Passer Rating

Wins piqued the highest level of interest for our analysis, because that is considered the ultimate success. However, each of the response variables mark some significant form of performance. QBNC and QBWC provide different perspectives for the linear model, so we selected our predictors accordingly. Table 1 below illustrates a total of 11 predictors for QBNC and 17 predictors for QBWC. Coupled with nine response variables, doing manual linear regression on each possible combination of predictors for a response variable proved futile. Consequently, we chose a method known as stepwise regression.

Table 1. Linear Regression Predictors

QBNC	QBWC
height	height
weight	weight
age	age
average completion percentage	average completion percentage
passer rating	passer rating
completions	completions
average number of interceptions	average number of interceptions
average number of touchdowns	average number of touchdowns
average number of yards	average number of yards
number of years in college	number of years in college
average number of attempts	average number of attempts
	40 yard dash
	Wonderlic score
	cone
	shuttle time

	vertical leap
	broad jump

3.1.1 Stepwise Regression

Our logical approach to linear regression was to find the best model from the set of predictors that minimizes the mean squared error. Stepwise regression allowed us to find this model in for a given set of predictors and the steps are as follows:

1. Start with empty model and specify significance level.
2. Fit response variable on Xi to Xp-1 and pick best Xi. (based on test)
3. Add best fit Xi from remaining Xs
4. Iteratively check that adding Xi does not reduce the previously added Xs
5. Stop when algorithm can't add another X best on test

This is known as bi-directional stepwise regression because the algorithm goes back and tests former predictors added to the model to see if they were affected by the new addition [6]. Despite the documented caveats associated with stepwise regression such as model complexity and a lack of human interaction with the model building, we chose to go forward with this method [7]. The following sections highlight the performance of stepwise regression performed on scaled, and log transformations on QBNC and QBWC.

3.1.2 Scaled Transformation

We normalized our data prior to running our first stepwise regression on our response variables, because we did not want the values of some of our predictors to heavily influence the model due to their magnitude. Statistics such as yards will be significantly higher than other stats such as wins or interceptions. Tables 2 and 3 in Appendix A provide an overview of our results for QBNC and QBWC after stepwise regression. Each table contains the response variable, final predictors, R-squared value, and residual standard error for the optimal model, as determined by the bi-directional stepwise regression. Not having the combine data clearly affected the

linear models with the highest R-Squared value for QBNC registering at 0.198 for completion percentage. However, there was only mild improvement when adding the combine data. Interceptions, completion percentage, and NFL wins registed R-Squared values of greater than 0.4. These results indicate that 40% of the variance in each of these response variables can be identified by their respective predictors. Figure 3 below provides an output of the cross validation performed on the interceptions model for QBWC and, though better than results found without combine data, indicate lackluster results at best with a mean square prediction error of 0.765.

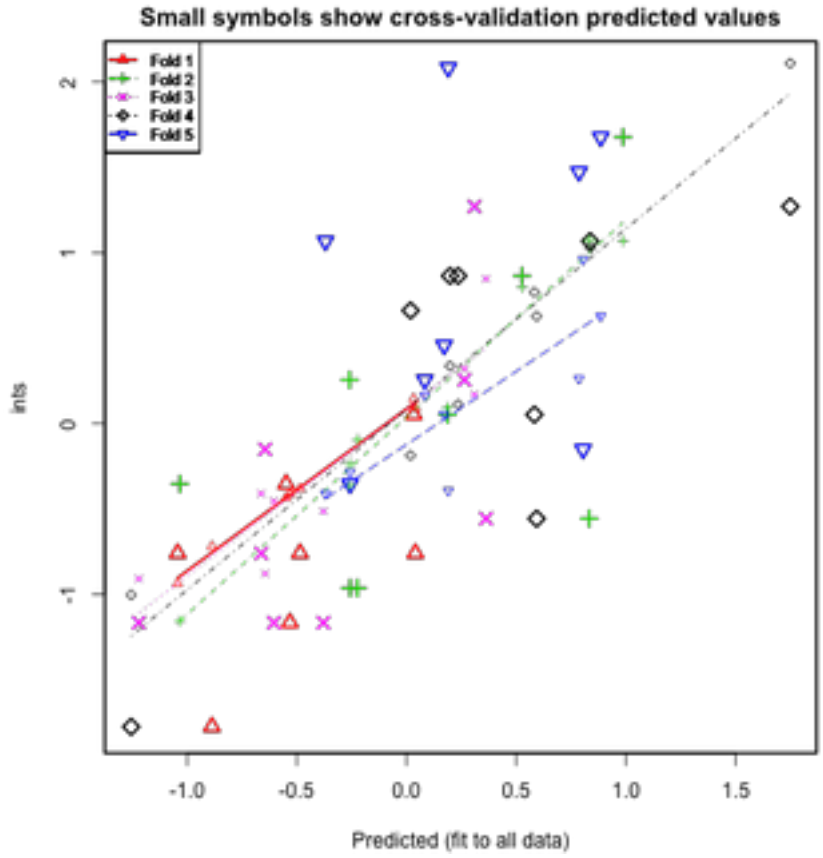

Figure 3. Cross Validation on Interceptions (QBWC)

3.1.3 Log Transformation

After seeing some tails in the histograms[12] for each predictor attribute, we decided to perform a log transformation on all the data.[13] Table 4 and Table 5 provide an overview of our results for QBNC and QBWC after stepwise regression. Similar to Table X and Y, the columns contain the response variable, predictors, R-Squared, and residual standard error. The residual standard errors were reduced across the board for the models including the log transformed data. However, there was minor improvement in the R-Squared values for each model which indicates the log transformation provided little benefit. The model for completion percentage provided the most interesting results. The cross validation error for this model was 0.0652. Figure 4 provides an output of the five fold cross validation.[14]

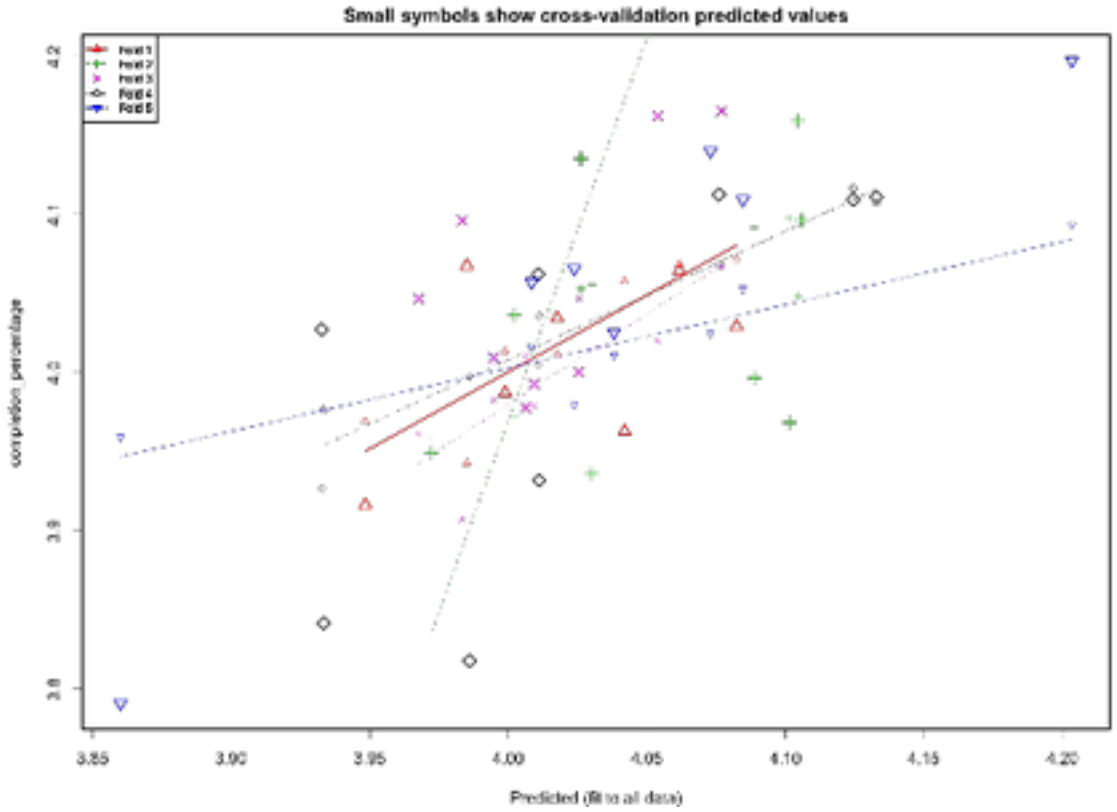

Figure 4. Cross Validation on Completion Percentage

3.2 LOGISTIC REGRESSION

[12] http://bit.ly/qb_histograms

[13] log(x + 0.1)

[14] For complete linear regression analysis visit http://bit.ly/qb_linreg

We utilized logistic regression for classification of certain types of quarterbacks based on their NFL statistics. In contrast to linear regression, logistic regression models proved more successful in identifying connections between the college and NFL statistics. The following sections cover our efforts using logistic regression as a classifier.

3.2.1 Binarization

Our initial data set did not lend itself well to any form of classification. All of our response variables are continuous so, the only way in which classification can be performed is through some form of binarization of response variables. However, choosing thresholds on which to binarize each attribute can be subjective and provide trivial prediction results. We chose to perform a feature extraction from a subset of response variables and use these new binary variables for logistic regression. The obvious target of interest in football is wins, because they are the only thing that get teams to the playoffs. However, teams are not required to win every game to get into the playoffs. When a team wins at least 50% of their games their chances of making the playoffs increase dramatically [8]. Consequently, we created a win percentage variable based on each quarterback's games started and games won. We then binarized the percentages on the threshold 0.5. If a quarterback won more than 50% of their games then they were assigned one for that variable and zero otherwise. We considered the quarterbacks assigned one significantly more likely to lead a team to the playoffs.[15]

NFL scouts have discovered some trends in the completion percentage of a quarterback. The general agreed threshold is 60%. If a quarterback can complete 60% of their passes then they are more likely to be successful in the NFL [9]. We selected our threshold to be 60% as well for the logistic regression. The final feature we created was the ratio between touchdowns and interceptions. In general, quarterbacks that throw a maximum number of touchdowns and a minimum number of interceptions are desired. However if two quarterbacks threw 25 touchdowns in their debut season, coaches are going to want the player who threw 12 interceptions instead of 20. In this scenario, it's the ratio of the touchdowns to interceptions that is interesting. We decided to binarize the new variable on a ratio of two to one, to ensure that the observation of one variable is always considered in the context of the other. If a quarterback

[15] Note that not every team that wins over 50% of games makes the playoffs.

throws at least twice as many touchdowns as interceptions they are assigned the value one. Otherwise, they are given the value zero.

3.2.2 Evaluating the Classification

We used the QBNC and QBWC with no modification to the predictors stated previously. However, we discovered that the lack of quarterback data in QBWC introduced convergence problems for the logistic regression model for each response variable. In both data sets the entire set was used as the training data. We found that we did not have enough data to provide explicit training and test data for the models. We found that with 8.7% error our model could predict whether a quarterback would throw more than twice as many touchdowns as interceptions. This proved to be the highlight of the logistic regression analysis.[16] Furthermore, Table 6 shows in sorted order the cross validation error of each of the models built with QBNC.

Table 6. Cross Validation Error for Logistic Regressions (QBNC)

Response	Cross Validation Error
Touchdown Interception Ratio	0.08658
Completion Percentage	0.1483
Win Percentage	0.4185

3.3 PCA

We performed Principal Component Analysis (PCA)[17] on the the college statistics to examine what features account for the most variance in the college stats data set. Some of the college stats that we gathered are actually compositions of other stats. For example, c_ya and c_aya are both based on yards and and attempts, both of which are already accounted for in the data. These stats were not considered when performing PCA, because they are highly correlated with the other raw stats in the data set.

Table 7 shows the result of PCA for the most important components:

Table 7. PCA Output

[16] For complete logistic regression analysis visit http://bit.ly/qb_logreg

[17] http://rpubs.com/sgscaffidi3/PCA

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	2.1896	1.3794	1.0171	0.97667
Proportion of Variance	0.4814	0.1911	0.1039	0.09579
Cumulative Proportion	0.4814	0.6725	0.7764	0.87218

Almost half of the variance can be explained by the first component, while the four components together explain about 88%. Table 8 lists the makeup of each of these four components:

Table 8. PCA Breakdown

	Comp.1	Comp.2	Comp.3	Comp.4
age	0.08522	0.06529	0.827855	0.483646
height	0.14417	0.47167	0.221787	0.528711
weight	0.2125	0.53291	0.202958	0.179713
c_numyrs	0.15297	0.31314	0.301515	0.633097
year	0.30288	0.36613	0.324327	0.075072
c_avg_cmpp	0.44017	0.12729	0.044527	0.015472
c_avg_att	0.43665	0.17139	0.016788	0.021336
c_avg_tds	0.41183	0.10894	0.000394	0.063305
c_avg_inter	0.25236	0.42335	0.157788	0.207418
c_avg_yds	0.4404	0.14313	0.031839	0.009049

Component 1 appears to capture football statistics, with the exception of interceptions. Component 2 accounts for those interceptions, but also captures the physical characteristics height and weight. Component 3 captures age and component 4 captures the number of years played in college. The year the player was drafted is almost equally represented in all of the first three components.

Attempting to use the principal components in linear regression did not result in any viable models[18], so the focus of the PCA effort went towards identifying trends in the data over time. The subset of data used for PCA was divided into 8 groups based on the decade that the player first started in the NFL. Figure 5 below shows the scores of the first two components, with the colors indicating the decade that the quarterback first started in the NFL.

[18] http://rpubs.com/sgscaffidi3/pcalm

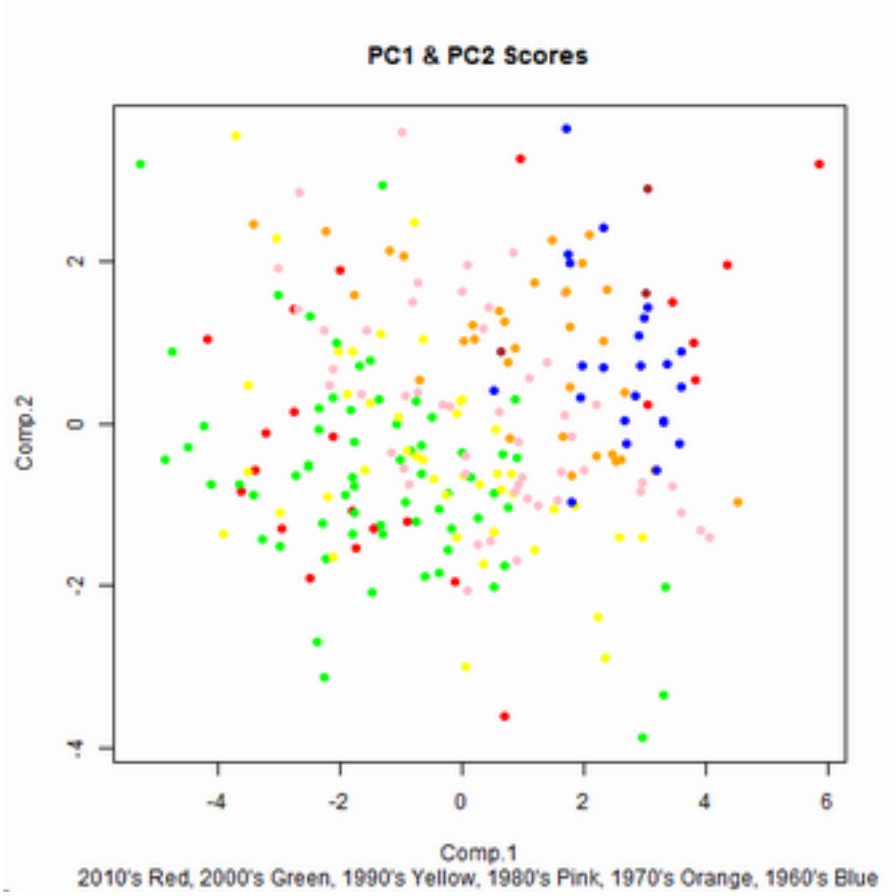

Figure 5. College Career Statistics by Decade

PCA could not be performed on each of these groups, because the 40's and 50's contained too few players. There was only one quarterback in the 1940s and three in the 1950s in the data set under analysis. Instead, PCA was performed on the 6 groups representing the 1960's -2010's, and the loadings for each one of the variables were plotted for each of the four major principal components over time.

Unfortunately, the results from this experiment did not lead to many interesting observations. The number of quarterbacks per decade has also steadily increased since the 1940s, and since the size of our data set is relatively small, this may have had a significant impact on the changes in variance we observed. From looking at PC1 over time, we can see the charts for average yards, attempts, and completions following a similar pattern. Component 2 shows that height remains fairly consistent from the 70's through today.

3.4 CLUSTERING

We applied two types of clustering to our data, two-dimensional clustering and multidimensional clustering using K-means. In our two-dimensional clustering analysis we applied several cluster sizes using K-means to individual college career statistics against individual nfl statistics. We analyzed clusters of size two, three, and four. Based on our prior knowledge of college and pro football we felt that these clustering sizes made the most sense, since our main objective was to differentiate the performance of broad groups of quarterbacks. In this case we felt like cluster groups would fall under a few general categories: low, average, and high performing quarterbacks. As a result we limited our cluster size to two, three, and four. Two-dimensional clustering yielded poor results. We ran ten individual college career predictors against thirty individual target variables from the NFL first year quarterback season with a data set of 240 quarterbacks.

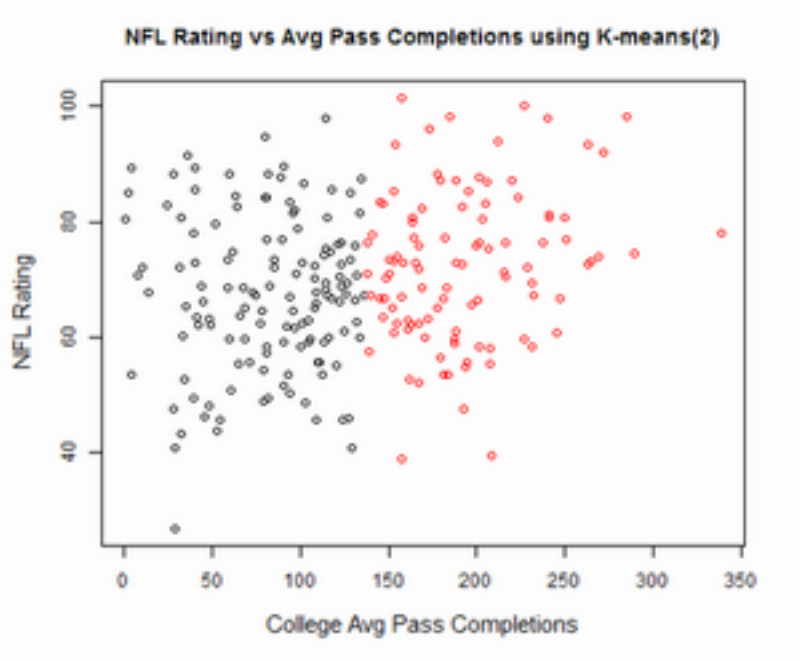

Figure 6. 2D Clustering NFL Passer Rating vs College Avg Pass Completions

An example of this is a K-means cluster of 2 on NFL passer rating vs Average Pass Completions in college[19]. The result is shown in Figure 6 above. We used the between sum of squares over the total sum of squares ratio as a guidance for evaluating each result. This particular cluster had

[19] http://rpubs.com/idfg/NFLRating

a ration of 65.5%. In general, the ratio ranged between 37% and 71%. The clusters did not yield low similarity within each cluster and high similarity between clusters. Additionally, since we were analyzing two-dimensional clusters we could visually assess the results and verify that K-means was not biasing our results.

Our second approach was to apply multidimensional clustering to the college career data and use the resulting clusters to identify potential trends in the NFL data. We ran multi-dimensional k-means of two, three, and four clusters. The resulting clusters were then used to cluster the NFL data into their respective cluster groups to evaluate characteristics of each clusters. We compared the mean of each individual target variable for each cluster group. This also yielded poor results, since in the first phase the clusters also had poor between sum of squares and total sum of square coefficients[20].

4.0 POST WORK LEARNING

Sports research is something new to us. So, looking back there were some areas of our project that could have been improved upon. This section highlights some of the difficulties associated with predicting quarterback statistics and some potential future endeavors in the vein of this same topic.

4.1 CHALLENGES

The lack of consistency across a set of attributes added difficulty to maintaining a large data set of quarterbacks. We were required to splice data from multiple sources into our data set and every time we added a new set of attributes our number of quarterbacks would reduce because one source would not have information about all the quarterbacks we had previously. Initially we scraped over 900 quarterbacks data from Pro Football Reference[21] and after we spliced all of our data together we dropped to as low as 54 quarterbacks that contained every attribute we wanted to examine. We were forced to make a compromises on which attributes we chose to incorporate into our models based on the sparsity and lack of consistency across our data sources.

[20] http://rpubs.com/idfg/MCKMeans2

[21] http://www.pro-football-reference.com/players/

Even with all of our data, the scope of our problem is inherently difficult. Football is different from sports like basketball where a superstar player will carry a team to success. There are many factors that contribute to how well a quarterback performs that range from the team they play on to the psychological difficulty of playing the position at such a high level. We tried to limit the scope of our research to statistics that pertain only to the quarterback but the fact remains that other players influence these numbers. The amount of sacks a quarterback has can be a strong reflection of how quickly they make their reads in the pocket or it could be directly correlated with how poor an offensive line they are behind. Similarly, some quarterbacks have a star running back that a team will run the offense through. Consequently, the team's offensive success requires less of this quarterback. This may influence statistics such as completion percentage because that quarterback does not have to throw as many passes. As it stands our research does not take into account these nuances of the attributes for each quarterback and it could have explained the difficulty finding answers in our data sets.

4.2 FUTURE WORK

Any future work would require a more complete data set to use with our models. We would have to explore other more consistent means of retrieving our data. Unfortunately this would probably require us to pay for our data. However, having a larger data set would allow us to train our models using some training data as opposed to having all of our data train the models we use. In addition, we should try to explore the career of an NFL quarterback as opposed to their debut season. This might yield clearer results as an aggregation of data over a period of time would provide more consistency in the response variables at the pro level.

5.0 CONCLUSION

The outcome of research highlights that the quarterback prediction problem remains a difficult one. However, some aspects of our research indicate some connection between a quarterback's college and NFL performance. We believe this notion can be improved upon and as more research unfolds about which NFL stats for a quarterback translate to wins. Knowing which stats

are critical at the NFL level can allow us to hone our research on the stats that contribute to wins as opposed to the actual wins themselves. Statistically identifying another Peyton Manning or Tom Brady may still be impossible but we believe this type of research can aid in mitigating the risk of wrongfully going all in on a quarterback who may yield pedestrian performance.

REFERENCES

[1] "1997 Heisman Trophy Voting." Internet: http://www.sports-reference.com/cfb/awards/heisman-1997.html, [April 15, 2013].

[2] Pullman. *CNNSI.* "Where will Leaf fall?" Internet: http://sportsillustrated.cnn.com/football/nfl/events/1998/nfldraft/news/1998/04/15/leaf_package/, [April 17, 2013].

[3] Anderson. *The Seattle Times.* "Manning's Deal A Rookie Record..." Internet: http://community.seattletimes.nwsource.com/archive/?date=19980729&slug=2763598, [April 15, 2013].

[4] "Peyton Manning leads...." Internet: http://www.nydailynews.com/sports/football/peyton-manning-leads-indianapolis-colts-9th-straight-playoff-appearance-host-jets-saturday-article-1.148277, [April 15, 2013].

[5] *Pro-Fooball-Reference.* "Ryan Leaf" Internet: http://www.pro-football-reference.com/players/L/LeafRy00.htm, [May 1, 2013].

[6] "Stepwise Regression." Internet: https://onlinecourses.science.psu.edu/stat501/node/88, [May 1, 2013].

[7] Flom, Peter. "Stopping stepwise: Why stepwise..." Internet: http://www.nesug.org/proceedings/nesug07/sa/sa07.pdf, [May 1, 2013].

[8] "Sports Club Stats." Internet: http://www.sportsclubstats.com/NFL.html, [May 2, 2013].

[9] Lopez, John. *SI.com.* "The Rule of 26-27-60 helps predict NFL..." Internet: http://sportsillustrated.cnn.com/2010/writers/john_lopez/07/08/qb.rule/index.html, [May 1, 2013].

APPENDIX A – LINEAR REGRESSION RESULTS

APPENDIX A – LINEAR REGRESSION RESULTS

Table 2. No Combine Data (Scaled)

Response	Predictors	R-Squared	Residual standard error
Completion Percentage	weight + age + c_avg_comp + c_numyrs + c_avg_att	0.198	0.905
Passer Rating	weight + age + c_avg_comp + c_avg_att	0.158	0.925
Yards	weight + age + c_avg_yds + c_avg_att	0.143	0.933
Games Started	weight + c_avg_yds + c_avg_att	0.0855	0.962
Interceptions	age + c_avg_inter + c_avg_yds + c_avg_att	0.0812	0.967
NFL Wins	weight + c_rate + c_avg_yds + c_avg_att	0.0606	0.0444
Touchdowns	weight + age + c_avg_cmp + c_avg_yds	0.0475	0.984
Sacks	c_avg_inter + c_avg_yds	0.0174	0.996

Table 3. With Combine Data (Scaled)

Response	Predictors	R-Squared	Residual standard error
Interceptions	age + X40 + cone + broad_jump	0.461	0.776
Completion Percentage	weight + age + c_avg_cmpp + c_avg_tds + c_avg_att + cone + shuttle + vert_leap	0.46	0.83
NFL Wins	age + c_rate + c_pct + c_avg_tds + c_avg_yds + c_numyrs + c_avg_att + X40 + wonderlic + broad_jump	0.444	0.869
Passer Rating	weight + age + c_avg_cmpp + X40 + vert_leap	0.337	0.876
Games Started	age + c_avg_cmpp + c_avg_att + wonderlic	0.317	0.873
Yards	height + age + c_avg_cmpp + c_avg_att + wonderlic	0.298	0.899
Sacks	c_avg_yds + c_avg_att	0.246	0.203
Touchdowns	age + c_avg_yds + wonderlic	0.144	0.965

Table 4. No Combine Data (Log)

Response	Predictors	R-Squared	Residual standard error
Passer Rating	height + weight + age + c_avg_cmpp + c_pct + c_avg_att	0.166	0.189
Interceptions	height + age + c_avg_inter + c_avg_yds + c_avg_att	0.09	0.517
Games Started	weight + c_avg_cmpp + c_rate + c_pct + c_avg_yds + c_avg_att	0.0849	0.263
Yards	weight + c_avg_yds + c_avg_att	0.0512	0.749
NFL Wins	c_avg_cmpp + c_pct + c_avg_tds + c_avg_yds + c_avg_att	0.0495	0.923
Sacks	c_avg_cmpp + c_rate + c_avg_inter + c_avg_tds + c_avg_yds + c_avg_att	0.0421	0.586
Touchdowns	c_avg_cmpp + c_avg_att	0.0148	0.698
Completion Percentage	c_avg_yds + c_avg_att	0.0147	0.421

Table 5. With Combine Data (Log)

Response	Predictors	R-Squared	Residual standard error
Passer Rating	weight + age + c_avg_cmpp + c_rate + c_pct + c_numyrs + c_avg_att + X40 + wonderlic + vert_leap	0.478	0.152
Completion Percentage	c_avg_cmpp + c_rate + c_pct + c_numyrs + c_avg_att + X40 + cone + vert_leap	0.464	0.0788
Interceptions	age + cone	0.436	0.367
Yards	height + c_avg_cmpp + c_rate + c_pct + c_avg_tds + c_avg_yds + c_numyrs + c_avg_att + wonderlic	0.326	0.379
Touchdowns	weight + c_avg_cmpp + c_pct + c_avg_inter + c_numyrs + c_avg_att + wonderlic	0.314	0.572
Games Started	age + c_avg_cmpp + c_avg_att + wonderlic	0.245	0.271
Sacks	c_avg_yds + c_avg_att	0.162	0.383
NFL Wins	c_avg_inter	0.0824	0.895